LECTURE 4 : GAUSSIAN DISTRIBUTION

[1] Motivation

Both MLE & MAP give us the prediction of $\bar{w}^*$ in the training stage : $\bar{w}^* = \underset{\{\bar{w}\}}{\operatorname{argmax}} \; p(\bar{w} | \bar{x}, \bar{t})$.

This $\bar{w}^*$ is later used in the testing stage:

$$t_{new} = \bar{w}^* \bar{x}_{new} \quad \text{or} \quad t_{new} = \bar{w}^* \phi(\bar{x}_{new})$$

Q: So, what's wrong with MLE or MAP?

A1: Both MLE & MAP provide the prediction ($\bar{w}^*$), but they don't provide any information on the probability or uncertainty information. Namely, how certain the predicted $\bar{w}^*$ is.

A2: This probability information is important in decision making in general, but also critical when we need to compare two or more prediction results. Of course, we will choose or be more confident to use the predicted result with the highest probability or certainty.

Full Bayesian Inference resolves the problem by using:

$$p(x|d) = \frac{p(d|x)\, p(x)}{p(d)}$$

where :

$$p(d) = \int p(d,x)\, dx \cong \sum_x p(d,x)$$

$$= \int p(d|x) p(x)\, dx \cong \sum_x p(d|x) p(x)$$

Q: To obtain $p(x|d)$ requires $p(d)$. Then, how to obtain $p(d)$?

A: $$p(d) = \int p(d|x)\,p(x)\,dx \approx \sum_x p(d|x)\,p(x)$$

There are many possible functions for $p(d|x)\,p(x)$

↓

Many of them are difficult to be integrated (analytically intractable).

e.g.: $\int \exp(x^3 + x^2)\,dx$

In discrete, it can be computationally expensive.

Q: Why is $\sum_x p(d|x)\,p(x)$ computationally prohibitively expensive?

A: Let's make $p(d) = \sum_{\bar{x}} p(d, \bar{x})$, where $\bar{x}$ is a vector.

If $\bar{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, then:

$$p(d) = \sum_{x_1}\sum_{x_2} p\left(d, \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \sum_{x_1}\sum_{x_2} p(d, x_1, x_2)$$

In programming it means:

```
for all candidates of (x1) do
    for all candidates of (x2) do
        p(d) = p(d) + p(d, x1, x2)
    end;
end;
```

} (# candidates)^2

Imagine if $\bar{x}$ is a vector with length of 100 & the number of candidates is 50, then the complexity is $(50)^{100}$

↓↓

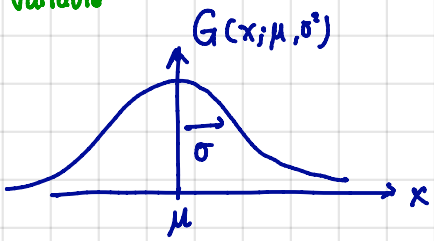
This is the reason why it is prohibitively expensive!

[2] Univariate & Multivariate Gaussian

Univariate:

$$G(x;\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

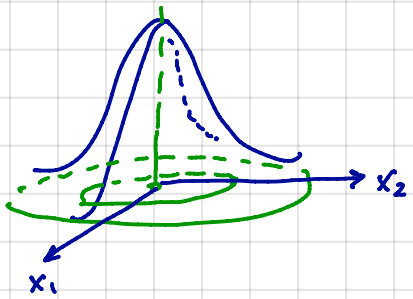
x: a scalar variable



Multivariate:

$$G(\bar{x};\bar{\mu},\Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\underset{1\times D}{\bar{x}-\bar{\mu}})^T \underset{D\times D}{\Sigma^{-1}} (\underset{D\times 1}{\bar{x}-\mu})\right)$$

$\bar{x}$ ($D\times 1$): a vector; Σ ($D\times D$): covariance matrix



Covariance Matrix, Σ:

Definition of covariance: $\text{cov}(a,b) = \mathbb{E}[(a-\mathbb{E}(a))(b-\mathbb{E}[b])]$ ($\mathbb{E}$: Expectation)

↓

Covariance Matrix

$$\bar{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix}$$

→ the covariance of x_i & x_j:

μ_i: the average of $(x_i^1, x_i^2, \ldots, x_i^N)$

$$\Sigma_{ij} = \text{cov}(x_i, x_j) = \mathbb{E}[(x_i-\mu_i)(x_j-\mu_j)]$$

Where: $\mu_i = \mathbb{E}[x_i]$

↓↓

Hence:

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \cdots & \Sigma_{1D} \\ \Sigma_{21} & \ddots & & \vdots \\ \vdots & & \ddots & \\ \Sigma_{2D} & \cdots & & \Sigma_{DD} \end{bmatrix}$$

$\text{cov}(a,b) = \mathbb{E}[ab] - \mathbb{E}[a]\mathbb{E}[b]$

↓

Question: What is the geometrical meaning of a covariance matrix?

[3] Covariance Matrix : Geometric Meaning

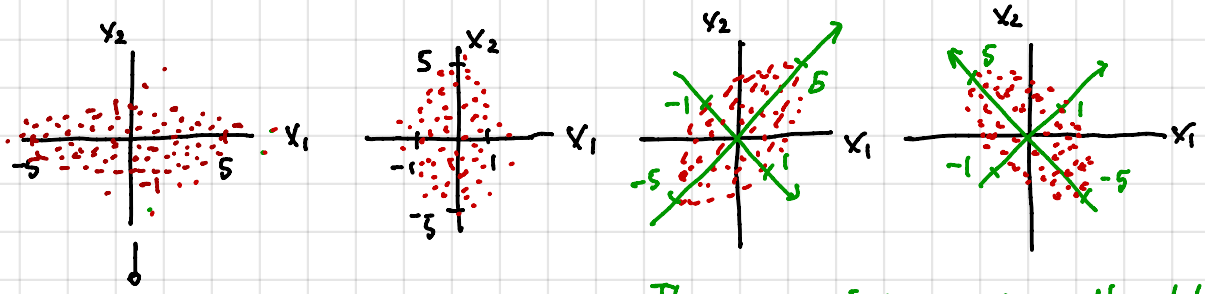
Let $\bar{x} \in \mathbb{R}^2$: $\Sigma_{2\times 2} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$ → a covariance matrix is always a symmetrix matrix :

$\Sigma_{21} = \Sigma_{12}$ ↓ $\Sigma^T = \Sigma$

Examples

Assuming $\bar{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$:

$\Sigma_A = \begin{bmatrix} 5 & 0 \\ 0 & 1 \end{bmatrix}$ $\Sigma_B = \begin{bmatrix} 1 & 0 \\ 0 & 5 \end{bmatrix}$ $\Sigma_C = \begin{bmatrix} 5 & 4 \\ 4 & 6 \end{bmatrix}$ $\Sigma_D = \begin{bmatrix} 5 & -4 \\ -4 & 6 \end{bmatrix}$



$\Sigma_{11} = 5$ meaning :

$cov(X_1, X_1) = E[(X_1 - \mu_1)^2]$

$= E[X_1^2] = 5$

The above points are supposedly obtained by random sampling from a Gaussian distribution with $\bar{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and Σ.

Q : How do we know the directions of the distributions ? → How can we draw the above figures? How can we justify those point distributions?

A : The directions have something to do with eigenvectors.

The definition of eigenvectors : $\Sigma \bar{v} = \lambda \bar{v}$

$\bar{v}$ = an eigenvector

λ = an eigen value

Q : How to compute λ and $\bar{v}$ from a given Σ ?

$A\bar{x} = 0$

$\Sigma \bar{v} - \lambda \bar{v} = 0$

$(\Sigma - \lambda I)\bar{v} = 0$ → Trivial solution (that we don't want): $\bar{v} = 0$

s.t. $\bar{v} \neq 0$ or $|\bar{v}| = 1$

Solution exists if : $\det(\Sigma - \lambda I) = 0$

$$\det \begin{pmatrix} \Sigma_{11} - \lambda & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} - \lambda \end{pmatrix} = 0$$

① $5x_1 + 6x_2 = 0$

$2x_1 + 3x_2 = 0$

$\begin{bmatrix} 5 & 6 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$

$A\bar{x} = 0$ → For this: $\bar{x}$ must be zero! → why? Try to solve the equations:

$5x_1 + 6x_2 = 0$

$2x_1 + 3x_2 = 0$

↓

the only solution is $x_1 = x_2 = 0$!

$x_1 = -6/5\, x_2$

$\frac{-12}{5} x_2 + 3x_2 = 0 \rightarrow x_2 = 0 \rightarrow x_1 = 0$

② $4x_1 + 6x_2 = 0$

$2x_1 + 3x_2 = 0$

$\begin{bmatrix} 4 & 6 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$ → even $\bar{x} \neq 0$ the equation holds

Hence: $\det(A) = 0$

Examples:

① $\Sigma = \begin{bmatrix} 5 & 0 \\ 0 & 1 \end{bmatrix}$ → Q: How do we know the distribution directions (or principal axes)?

Answer:

$$\det\begin{pmatrix} 5-\lambda & 0 \\ 0 & 1-\lambda \end{pmatrix} = 0 \rightarrow (5-\lambda)(1-\lambda) = 0$$

$$\lambda_1 = 5, \quad \lambda_2 = 1$$

For $\bar{v}_1$: $\Sigma \bar{v}_1 = \lambda_1 \bar{v}_1$

$$\begin{bmatrix} 5 & 0 \\ 0 & 1 \end{bmatrix}\bar{v}_1 = 5\bar{v}_1 \rightarrow \begin{bmatrix} 5 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} v_1^a \\ v_1^b \end{bmatrix} = \begin{bmatrix} 5v_1^a \\ 5v_1^b \end{bmatrix}$$

$$\begin{bmatrix} 5v_1^a \\ v_1^b \end{bmatrix} = \begin{bmatrix} 5v_1^a \\ 5v_1^b \end{bmatrix} \rightarrow \begin{matrix} v_1^a = 1 \\ v_1^b = 0 \end{matrix}$$

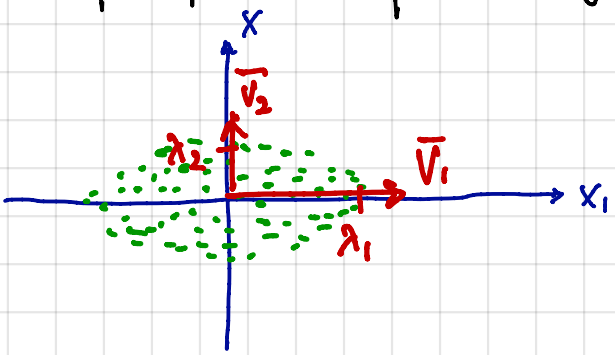
Thus: $\bar{v}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ ←

For $\bar{v}_2$: $\Sigma \bar{v}_2 = \lambda_2 \bar{v}_2$

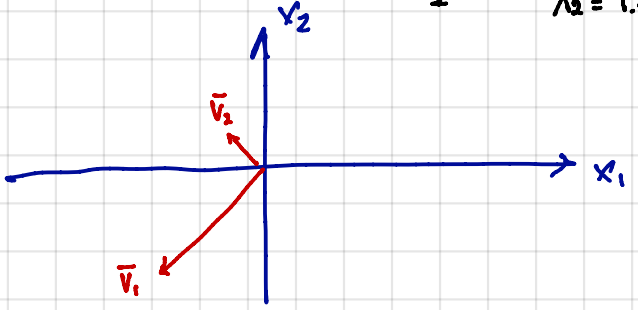
$$\begin{bmatrix} 5 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} v_2^a \\ v_2^b \end{bmatrix} = \begin{bmatrix} v_2^a \\ v_2^b \end{bmatrix} \rightarrow \begin{bmatrix} 5v_2^a \\ v_2^b \end{bmatrix} = \begin{bmatrix} v_2^a \\ v_2^b \end{bmatrix} \rightarrow \begin{matrix} v_1^a = 0 \\ v_2^b = 1 \end{matrix}$$

Thus: $\bar{v}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$

Hence now we know the principle axes represented by $\bar{v}_1$ & $\bar{v}_2$:



② $\Sigma = \begin{bmatrix} 5 & 4 \\ 4 & 6 \end{bmatrix}$ → Using the same technique: $\lambda_1 = 9.53$, $\lambda_2 = 1.47$ → $\bar{v}_1 = (-0.66, -0.75)^T$, $\bar{v}_2 = (-0.75, 0.66)^T$



CONDITIONAL & MARGINAL GAUSSIAN DIST.'s

Conditional probability distribution: $p(x|d)$

where: $p(x|d) = \frac{p(x,d)}{p(d)} = \frac{p(d|x)\,p(x)}{p(d)}$

Marginal distribution: $p(d)$

where: $p(d) = \int p(x,d)\,dx = \sum_x p(x,d)$

Motivation: we know that solving $p(d)$ and thus $p(x|d)$ is analytically intractable & computationally prohibitively expensive.

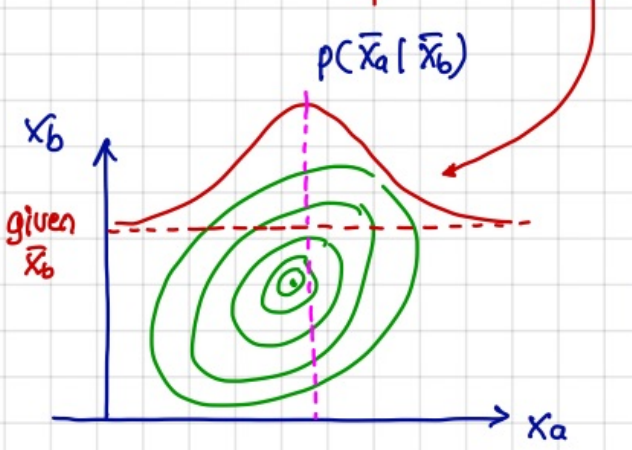
↓

Solution: to use the Gaussian distribution function

Two Cases

case 1:

We want to know the mean & the covariance matrix



We want to know:

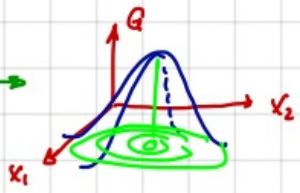
1. $p(\bar{x}_a|\bar{x}_b)$
2. $p(\bar{x}_a)$

case 2:

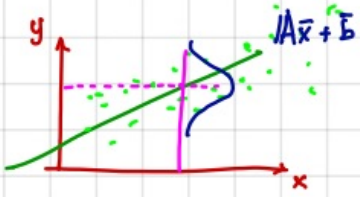
$\bar{y} = \mathbb{A}\bar{x} + \bar{b}$

D×1 D×M M×1 D×1

$p(\bar{x}) = G(\bar{x}; \bar{\mu}, \Lambda^{-1})$



$p(\bar{y}|\bar{x}) = G(\bar{y}; \mathbb{A}\bar{x} + \bar{b}, L^{-1})$



We want to know:

1. $p(\bar{x}|\bar{y})$
2. $p(\bar{y})$

CASE 1: $\bar{x} = \begin{pmatrix} \bar{x}_a \\ \bar{x}_b \end{pmatrix}$ and $p(\bar{x}) = G(\bar{x}; \mu, \Sigma)$

$$= G\left(\begin{bmatrix} \bar{x}_a \\ \bar{x}_b \end{bmatrix}; \begin{bmatrix} \bar{\mu}_a \\ \bar{\mu}_b \end{bmatrix}, \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix} \right)$$

Marginal Gaussian distribution:

$$p(\bar{x}_a) = G(\bar{x}_a; \mu_a, \Sigma_{aa})$$
$$p(\bar{x}_b) = G(\bar{x}_b; \mu_b, \Sigma_{bb})$$

Conditional Gaussian distribution:

$$p(\bar{x}_a | \bar{x}_b) = G(\bar{x}_a; \mu_{a|b}, \Sigma_{a|b}^{-1})$$

where: $\mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b)$

$\Sigma_{a|b} = (\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1}$

$$p(\bar{x}_b | x_a) = G(\bar{x}_b; \mu_{b|a}, \Sigma_{b|a}^{-1})$$

where: $\mu_{b|a} = \mu_b + \Sigma_{ba}\Sigma_{aa}^{-1}(x_a - \mu_a)$

$\Sigma_{b|a} = (\Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab})^{-1}$

CASE 2: $\bar{y} = A\bar{x} + \bar{b}$

where:

$$p(\bar{x}) = G(\bar{x}; \bar{\mu}, \Lambda^{-1}) \quad \rightarrow \text{Prior}$$
$$p(\bar{y}|\bar{x}) = G(\bar{y}; A\bar{x} + b, L^{-1}) \quad \rightarrow \text{Likelihood}$$

Marginal Gaussian distribution:

$$p(\bar{y}) = G(\bar{y}; A\bar{\mu} + \bar{b}, L^{-1} + A\Lambda^{-1}A^T)$$

Conditional Gaussian distribution:

$$p(\bar{x}|\bar{y}) = G(\bar{x}; \bar{\mu}_{x|y}, \Sigma_{x|y})$$

where:

$$\mu_{x|y} = (\Lambda + A^T L A)^{-1}(A^T L(\bar{y} - \bar{b}) + \Lambda\bar{\mu})$$
$$\Sigma_{x|y} = (\Lambda + A^T L A)^{-1}$$

• FULL BAYESIAN INFERENCE FOR REGRESSION

In the training stage, we need to estimate: $\bar{w}$

According to Bayes' rule:

$$p(\bar{w}, \bar{x} \mid \bar{t}) = \frac{p(\bar{t} \mid \bar{w}, \bar{x})\, p(\bar{w} \mid \bar{x})\, p(\bar{x})}{p(\bar{t})}$$

posterior

; we don't have any prior information on $\bar{x}$

$$= \frac{p(\bar{t} \mid \bar{w})\, p(\bar{w})}{p(\bar{t})}$$

likelihood, prior

We know:

- Likelihood: $p(\bar{t} \mid \bar{w}) = G(\bar{t}; \Phi\bar{w}, \beta^{-1}\mathbb{I})$
- Prior: $p(\bar{w}) = G(\bar{w}; 0, \alpha^{-1}\mathbb{I})$

} we want to know $p(\bar{w}, \bar{x} \mid \bar{t})$ or $p(\bar{w} \mid \bar{t})$

This problem fits well to CASE 2:

CASE 2: $\bar{y} = \mathbb{A}\bar{x} + \bar{b}$

where:

$p(\bar{x}) = G(\bar{x}; \bar{\mu}, \Lambda^{-1})$ → Prior

$p(\bar{y} \mid \bar{x}) = G(\bar{y}; \mathbb{A}\bar{x} + b, L^{-1})$ → Likelihood

Marginal Gaussian distribution:

$$p(\bar{y}) = G(\bar{y}; \mathbb{A}\bar{\mu} + \bar{b}, L^{-1} + \mathbb{A}\Lambda^{-1}\mathbb{A}^T)$$

Conditional Gaussian distribution:

$$p(\bar{x} \mid \bar{y}) = G(\bar{x}; \bar{\mu}_{x|y}, \Sigma_{x|y})$$

where:

$$\mu_{x|y} = (\Lambda + \mathbb{A}^T L \mathbb{A})^{-1}(\mathbb{A}^T L(\bar{y} - \bar{b}) + \Lambda\bar{\mu})$$

$$\Sigma_{x|y} = (\Lambda + \mathbb{A}^T L \mathbb{A})^{-1}$$

Where:

μ	→ 0	$\bar{y}$	→ $\bar{t}$
Λ^{-1}	→ $\alpha^{-1}\mathbb{I}$	$\bar{b}$	→ 0
$\mathbb{A}$	→ Φ	L	→ $\beta\mathbb{I}$
$\bar{x}$	→ $\bar{w}$		

Hence:

$$p(\bar{w} \mid \bar{t}) = G(\bar{w}; \bar{m}_N, \mathbb{S}_N)$$

$M \times 1$ $M \times 1$ $M \times M$

where:

$$\mathbb{S}_N = (\alpha\mathbb{I} + \Phi^T\beta\mathbb{I}\Phi)^{-1} = (\alpha\mathbb{I} + \beta\Phi^T\Phi)^{-1}$$

$$\bar{m}_N = \mathbb{S}_N(\Phi^T\beta\mathbb{I}\Phi + \alpha\mathbb{I}\phi) = \mathbb{S}_N\beta\Phi^T\bar{t}$$

Hence:

$$\underset{M\times 1}{\bar{m}_N} = \beta \underset{M\times M}{S_N} \underset{M\times N}{\Phi^T} \underset{N\times 1}{\bar{t}}$$

$$\underset{M\times M}{S_N} = (\underset{M\times M}{\alpha I} + \beta \underset{M\times M}{\Phi^T\Phi})^{-1}$$

In other words:

$$\bar{m}_N = \beta(\alpha I + \beta \Phi^T\Phi)^{-1}\Phi^T\bar{t}$$

MLE:

$$\bar{w}_{MLE} = (\Phi^T\Phi)^{-1}\Phi^T\bar{t}$$

$$\bar{w}_{MAP} = \beta(\alpha I + \beta\Phi\Phi)^{-1}\Phi t$$

they are the same!

Note: Unlike MAP, full Bayesian provides S_N which in turn gives us: $p(\bar{w}|\bar{t})$. This means we can obtain uncertainty.

• SEQUENTIAL LEARNING

[1] Sequential Bayesian Learning

Concept:

simplification:

$$p(w \mid t_1, t_2, t_3, x_1, x_2, x_3) = p(w \mid t_1, t_2, t_3)$$

→ we want to estimate $\bar{w}$, the curve parameters.

$$= \frac{p(t_3 \mid t_2, t_1, \bar{w})\; p(t_2, t_1, \bar{w})}{p(t_3, t_2, t_1)}$$

$$= \frac{p(t_3 \mid \bar{w})}{p(t_3)} \; \frac{p(t_2 \mid t_1, \bar{w})}{p(t_2)} \; p(t_1, \bar{w})$$

$$= \frac{p(t_3 \mid \bar{w})}{p(t_3)} \; \frac{p(t_2 \mid \bar{w})}{p(t_2)} \; \frac{p(t_1 \mid \bar{w})}{p(t_1)} \; p(\bar{w})$$

For illustration, see the textbook page #155.

Q: What is the implication of the previous formulation?

A: $$p(\bar{w} \mid t_1, t_2, t_3) = \frac{p(t_3 \mid \bar{w})}{p(t_3)} \; \frac{p(t_2 \mid \bar{w})}{p(t_2)} \; \frac{p(t_1 \mid \bar{w})}{p(t_1)} \; p(\bar{w})$$

likelihood 3 / evidence 3 — likelihood 2 / evidence 2 — likelihood 1 / evidence 1 — prior

posterior 1 : $p(\bar{w} \mid t_1) = \frac{p(t_1 \mid w)\, p(w)}{p(t_1)}$

Prior 2 ← posterior 2 : $p(\bar{w} \mid t_2, t_1) = \frac{p(t_2 \mid \bar{w})}{p(t_2)} \; p(\bar{w} \mid t_1)$

prior 3 ← posterior 3 : $p(\bar{w} \mid t_3, t_2, t_1) = \frac{p(t_3 \mid \bar{w})}{p(t_3)} \; p(\bar{w} \mid t_2, t_1)$

All these imply sequential learning or online learning, since we can process once we have a new data, and update the old stored value of $\bar{m}_N$ and S_N.

Q: Practically how can we implement it?

A: Like the discussion above, we need to define:

① Gaussian prior:

$$p(\bar{w}) = G(\bar{w}; 0, \alpha^{-1}\mathbb{I}); \quad \mu = 0, \ \Lambda = \alpha\mathbb{I}$$

② Gaussian Likelihood:

$$p(t_n|\bar{w}) = G(t_n; \bar{w}^T\bar{\phi}(x_n), \beta^{-1})$$

Once t_1 arrives, we compute: $p(\bar{w}|t_1) = G(\bar{w}; \bar{m}_1, S_1)$ → See the definitions of $\bar{m}_1$ & S_1 in the previous note.

Once t_2 arrives, we compute: $p(\bar{w}|t_2) = G(\bar{w}; \bar{m}_2, S_2)$

But, what are $\bar{m}_2, S_2$ here?

Q: $p(\bar{w}|t_2) = G(\bar{w}; \bar{m}_2, S_2)$ → how to compute this?

A: We know: $p(w|t_2) = \dfrac{p(t_2|\bar{w})\, p(\bar{w}|t_1)}{p(t_2)}$

where: $p(\bar{w}|t_1) = G(\bar{w}; \bar{m}_1, S_1)$; $p(t_2|w) = G(t_2; \bar{w}^T\bar{\phi}(x_2), \beta^{-1})$

$\bar{\phi}_2 = \begin{bmatrix} \bar{\phi}^T(\bar{x}_2) \end{bmatrix}$ (1×M; M×1)

case 2 | Hence:

$$p(\bar{w}|t_2) = G(\bar{w}; \bar{m}_2, S_2)$$

with: (see textbook page 153 eq.(3.50))

$$S_2^{-1} = S_1^{-1} + \beta \underset{M\times 1}{\bar{\Phi}_2^T} \underset{1\times M}{\bar{\Phi}_2}$$

$$\bar{m}_2 = S_2\left(S_1^{-1}\underset{M\times M}{\bar{m}_1} + \beta \underset{M\times 1}{\bar{\Phi}_2^T} t_2\right)$$

Q: How can we transform the above formulation to an algorithm?

A: We simply use the definitions of S_2^{-1} and $\bar{m}_2$ above sequentially:

$$S_n^{-1} = S_{n-1}^{-1} + \beta \bar{\Phi}_n^T \bar{\Phi}_n$$

$$\bar{m}_n = S_n\left(S_{n-1}^{-1}\bar{m}_{n-1} + \beta \bar{\Phi}_n^T t_n\right)$$

Note: $p(\bar{w}|t_1) = G(\bar{w}; \bar{m}_1, S_1)$; $p(t_2|w) = G(t_2; \bar{w}^T\bar{\phi}(x_2), \beta^{-1})$

Gaussian prior — Gaussian likelihood

CASE 2: $\bar{y} = \mathbb{A}\bar{x} + \bar{b}$

where:

$p(\bar{x}) = G(\bar{x}; \bar{\mu}, \Lambda^{-1})$ → Prior

$p(\bar{y}|\bar{x}) = G(\bar{y}; \mathbb{A}\bar{x}+b, L^{-1})$ → Likelihood

Marginal Gaussian distribution:

$$p(\bar{y}) = G(\bar{y}; \mathbb{A}\bar{\mu}+\bar{b}, L^{-1} + \mathbb{A}\Lambda^{-1}\mathbb{A}^T)$$

Conditional Gaussian distribution:

$$p(\bar{x}|\bar{y}) = G(\bar{x}; \bar{\mu}_{x|y}, \Sigma_{x|y})$$

where:

$$\mu_{x|y} = (\Lambda + \mathbb{A}^T L \mathbb{A})^{-1}(\mathbb{A}^T L(\bar{y}-\bar{b}) + \Lambda\bar{\mu})$$

$$\Sigma_{x|y} = (\Lambda + \mathbb{A}^T L \mathbb{A})^{-1}$$

$\bar{x} \to \bar{w}$

$\bar{y} \to t_2$

$\Lambda \to S_1^{-1}$

$\mathbb{A} \to \bar{\phi}(\bar{x}_2) \to \Phi_2$; $\bar{\phi}(\bar{x}_2)$: $M \times 1$, Φ_2: $N \times M$; $N = 1$

$L \to \beta$

$b \to 0$

$\mu \to m_1$

Hence: Posterior:

$$p(\bar{w}|t_2) = G(\bar{w}; \bar{m}_2, S_2)$$

($\bar{w}$: $M \times M$, $\bar{m}_2$: $M \times 1$, S_2: $M \times M$)

where:

$$S_2^{-1} = (S_1^{-1} + \Phi_2^T \beta \Phi_2)^{-1}$$

(S_2^{-1}: $M \times M$, S_1^{-1}: $M \times M$, Φ_2^T: $M \times 1$, β: 1×1, Φ_2: $1 \times M$)

$$\bar{m}_2 = S_2^{-1}(\Phi_2^T \beta t_2)$$

($\bar{m}_2$: $M \times 1$, S_2^{-1}: $M \times M$, Φ_2^T: $M \times 1$, t_2: 1×1)

PREDICTIVE DISTRIBUTION

Motivation: Instead of $\bar{w}$ and its probability, $p(\bar{w}|\bar{t})$, we are more interested in t_{new} and its probability, $p(t_{new}|\bar{t})$. This $p(t_{new}|\bar{t})$ is not obtained in the previous discussion.

$t_* = t_{new} = y_* + \varepsilon = \bar{w}^T \bar{x}_* + \varepsilon$:

$$p(t_* | x_*, \bar{t}, \bar{x}, \alpha, \beta) \equiv p(t_*|\bar{t}) = \int p(t_*, \bar{w}|\bar{t})\, d\bar{w}$$

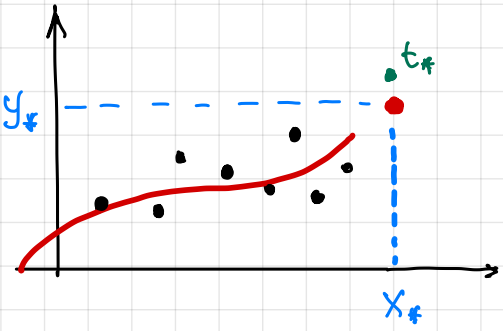
(t_*: 1×1, x_*: 1×1, $\bar{t}$: N×1, $\bar{x}$: N×1)

Marginalization over all possible values of $\bar{w}$.

$$= \int \underbrace{p(t_*|\bar{w})}_{①}\ \underbrace{p(\bar{w}|\bar{t})}_{②}\, d\bar{w}$$

① $p(t_*|\bar{w}) = G(t_*; \bar{w}^T \bar{\phi}(x_*), \beta^{-1})$

② $p(\bar{w}|\bar{t}) = \dfrac{p(\bar{t}|\bar{w})\, p(\bar{w})}{p(\bar{t})}$



where:

$$p(\bar{w}) = G(\bar{w}; 0, \alpha^{-1} I)$$

(M×1, M×1, M×M)

$$p(\bar{t}|\bar{w}) = G(\bar{t}; \Phi \bar{w}, \beta^{-1} I)$$

(N×1, N×1, N×N)

According to Bayesian sequential learning:

$$p(\bar{w}|\bar{t}, \alpha, \beta) = G(\bar{w}; \bar{m}_n, S_n)$$

where:

$$S_n^{-1} = S_{n-1}^{-1} + \beta \Phi^T \Phi$$

$$\bar{m}_n = S_n (S_{n-1}^{-1} \bar{m}_{n-1} + \beta \Phi^T \bar{t})$$

$\bar{m}_0 = 0$, $S_0^{-1} = \alpha$

If the prior and likelihood are Gaussians:

$$p(t_* \mid \bar{t}) = \int p(t_* \mid \bar{w})\; p(\bar{w} \mid \bar{t})\; d\bar{w}$$

① Gaussian Likelihood (w.r.t. $\bar{w}$)

② Gaussian Prior (w.r.t. $\bar{w}$)

then, according to case 2, the marginal distribution is:

case 2:

Recall the definition of the Marginal Gaussian distribution (lecture note 7):

Given: $p(x) = G(x; \mu, \Lambda^{-1})$ (prior)

$p(y|x) = G(y; Ax+b, L^{-1})$ (likelihood)

Then: $p(y) = G(y; A\mu+b, L^{-1} + A\Lambda^{-1}A^T)$

$$p(t_* \mid x_*, \bar{t}, \alpha, \beta) = G(t_*; \bar{m}_*, s_*)$$

(1×1; 1×1; 1×1)

$$\bar{m}_* = \bar{m}_N^T \bar{\phi}(x_*)$$

(1×1 = 1×M · M×1)

$$s_* = \beta^{-1} + \bar{\phi}^T(x_*) S_N \phi(x_*)$$

(1×1 = 1×M · M×M · M×1)

There are 2 steps: ① Prediction ② Update

See textbook page #156.

This implies that given the training data $\{(x_1, t_1) \dots (x_n, t_n)\}$ we can intrapolate or extrapolate the prediction of t and know the uncertainty of the prediction!

see textbook figure 3.8 (page #157)

If there are two data $[t_{*1}, t_{*2}]$ we want to process at once:

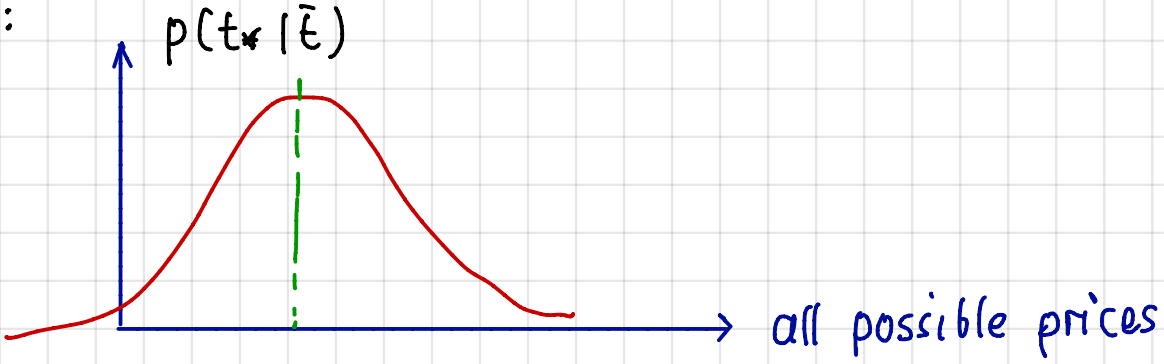
$$\bar{m}_* = \Phi_{1,2}\, \bar{m}_N \quad ; \quad S_* = \beta^{-1} I + \Phi_{1,2}^T S_N \Phi_{1,2}$$

(2×1 = 2×M · M×1; 2×2 = 2×2 + 2×M · M×M · M×2)

Note:

Q: Why is $p(t_* | \bar{t})$ beneficial?

A: If we know $p(t_* | \bar{t}) = G(t_*; \bar{m}_*, s_*)$, it means we obtain:



This means: we know the probability of all possible prices, and we know the most probable price.

→ (prediction & its uncertainty)

Q: How can we know that: $p(t_* | \bar{t}) = \int p(t_*, \bar{w} | \bar{t}) \, d\bar{w}$?

A: Basic idea of $p(t_* | \bar{t})$ is that we need to find the correlation between t_* and $\bar{t}$.

We know that in the testing, to estimate t_* we need x_* and $\bar{w}$ to obtain t_*; Also, in the training, we can get $\bar{w}$ from $\bar{t}$. Hence, the correlation of t_* & $\bar{t}$ can be obtained through $\bar{w}$. Yet, at the same time we don't care about $\bar{w}$ in the end. Therefore, we marginalize over $\bar{w}$ to obtain $p(t_* | \bar{t})$.